Software News and Update Reconstruction of Atomistic Details from Coarse-Grained Structures

Rzepiela, Andrzej J.; Schafer, Lars V.; Goga, Nicolae; Risselada, H. Jelger; De Vries, Alex H.; Marrink, Siewert J.

Published in:
Journal of Computational Chemistry

DOI:
10.1002/jcc.21415

Publication date:
2010

Software News and Update
Reconstruction of Atomistic Details from Coarse-Grained Structures

ANDRZEJ J. RZEPIELA, LARS V. SCHÄFER, NICOLAE GOGA, H. JELGER RISSELADA, ALEX H. DE VRIES, SIEWERT J. MARRINK

Groningen Biomolecular Sciences and Biotechnology Institute & Zernike Institute for Advanced Materials, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands

Received 18 May 2009; Revised 12 August 2009; Accepted 14 August 2009
DOI 10.1002/jcc.21415
Published online 19 January 2010 in Wiley InterScience (www.interscience.wiley.com).

Abstract: We present an algorithm to reconstruct atomistic structures from their corresponding coarse-grained (CG) representations and its implementation into the freely available molecular dynamics (MD) program package GROMACS. The central part of the algorithm is a simulated annealing MD simulation in which the CG and atomistic structures are coupled via restraints. A number of examples demonstrate the application of the reconstruction procedure to obtain low-energy atomistic structural ensembles from their CG counterparts. We reconstructed individual molecules *in vacuo* (NCQ tripeptide, dipalmitoylphosphatidylcholine, and cholesterol), bulk water, and a WALP transmembrane peptide embedded in a solvated lipid bilayer. The first examples serve to optimize the parameters for the reconstruction procedure, whereas the latter examples illustrate the applicability to condensed-phase biomolecular systems.

 J Comput Chem 31: 1333–1343, 2010

Key words: back-mapping; multiscale simulation; coarse-grained; molecular dynamics

Introduction

Molecular dynamics (MD) simulations have been successfully used to investigate biomolecular systems, such as proteins, DNA, or lipid membranes.[1–3] The huge computational effort involved in conventional atomistic MD simulations currently limits accessible simulation times to hundreds of nanoseconds and length scales to tens of nanometers. However, most biomolecular processes occur at slower time scales and, at the same time, often involve larger length scales. To study such processes, coarse-grained (CG) models, in which several atomistic particles are grouped together into effective beads, have recently gained more and more popularity in the field.[4] The main benefit of these models is their computational efficiency due to the reduced number of interaction sites. In this way, many CG methods allow probing the structural dynamics of large systems on time scales up to milliseconds and length scales up to hundreds of nanometers. This large gain in efficiency, however, comes at the cost of a reduced accuracy compared to atomistic (or all-atom, AA) models due to the inherent simplifications. Thus, tools are desirable that allow switching between the different levels of resolution, i.e., to back-map the atomistic structural ensemble that underlies a CG representation, thereby combining the efficiency of CG with the accuracy of AA models. This would allow CG models to live up to their full potential by applying them to explore large regions of phase space, followed by zooming in on the atomistic details of interesting configurations.

In this contribution, we present our recent implementation of an algorithm to reconstruct AA structures from their corresponding CG representations into the fast and freely available MD program package GROMACS.[5,6] The basic idea of the reconstruction algorithm is to carry out a simulated annealing (SA)[7] MD simulation of an atomistic system that is coupled to its corresponding CG system via restraints. During the SA, the system is cooled down from a high initial temperature to a desired target temperature, thus allowing the system to cross energy barriers and optimize under the restraints. Finally, the coupling is gradually removed to ensure a smooth relaxation of the final AA structure.

By means of a number of example applications presented in this work, we systematically and extensively tested the reconstruction procedure for different molecule types. Proper reconstructed ensembles were obtained using optimized parameters. Furthermore, although the applications presented in this contribution are based on the MARTINI CG force field[8–10] and the GROMOS96 AA force field,[11,12] the algorithm is general and can be applied in a

Correspondence to: S.J. Marrink; e-mail: s.j.marrink@rug.nl

Contract/grant sponsor: Netherlands Organisation for Scientific Research (NWO), Veni grant; Contract/grant number: 700.57.404

Contract/grant sponsor: National Supercomputing Facilities (NCF); Contract/grant numbers: NRG.2008.04, SH.002.08

straightforward manner to a wide range of force fields and molecule types, because it does not rely on libraries of predefined fragments.

A number of techniques to reconstruct atomistic details from CG structures have been proposed in the literature. Some of these methods[13–15] make use of a multiscale Hamiltonian exchange approach, in which the system is at the same time represented at both levels of resolution (and at several intermediate, i.e., mixed levels), and continuous attempts are made along an MD trajectory to switch between neighboring representations. The adaptive resolution schemes use a spatial compartmentalization of the simulation system and allow an on-the-fly interchange between atomistic and CG representations.[16–20] Other methods are more similar to the approach presented here in that they aim at reconstructing atomistic structures from single CG structures.[21–27] Most of these methods differ from the approach presented in this work by using libraries of predefined fragments to construct the initial atomistic structure, followed by a minimization and equilibration protocol. Additionally, these methods were often optimized only for the specific application at hand and may therefore not be generally applicable and transferable to other molecular systems and force fields.

The rest of this article is organized as follows. We set out to describe the methodological basis of the algorithm and detail the implementation. Additionally, we present details of the user interface. Subsequently, a number of examples will serve to illustrate the application of the algorithm, optimization of SA parameters, and generation of unbiased distributions. As a first example, the reconstruction algorithm was applied to obtain atomistic structural ensembles of different biomolecules in vacuum: the NCQ tripeptide, the dipalmitoylphosphatidylcholine (DPPC) lipid molecule, and cholesterol. Then, a box of atomistic water was reconstructed from CG water. Finally, a system composed of the WALP20 transmembrane peptide embedded in a solvated DPPC bilayer was reconstructed from its CG representation. The latter examples demonstrate the suitability of the reconstruction algorithm to condensed-phase biomolecular systems. A short conclusive section ends this article. In the appendix, we give a list of GROMACS commands used to carry out the reconstruction simulations.

Methods

Our goal is to generate a low-energy atomistic (AA) ensemble that underlies its corresponding CG system. This is achieved by means of a three-step reconstruction algorithm. Initially, AA particles are positioned close to their reference CG beads. Then, a SA procedure is used, during which the AA system is coupled to the CG system via harmonic restraints. Finally, the restraints are gradually removed to yield a relaxed atomistic system. As a prerequisite, the reconstruction algorithm requires the definition of a mapping of the AA structure to the CG structure, i.e., a prescription of which AA particles are represented by which CG beads. The mapping can be, e.g., defined via the center of mass of AA particles (as done in this work, see later) or via the positions of the C_α atoms in an amino acid chain.

Initial Placement of Atomistic Particles

To generate a starting configuration, the AA particles are randomly positioned within a sphere around their corresponding CG beads. In the MARTINI CG force field, on average four heavy atoms are mapped onto one CG bead. For the applications based on the MARTINI force field presented in this work, the radius of the sphere was chosen as $r_{CG} = 0.3\,\text{nm}$, which roughly corresponds to the typical van der Waals radius of a MARTINI bead. Such an initial placement of AA particles close to their expected final positions significantly speeds up the reconstruction procedure for condensed-phase systems.

Restrained Simulated Annealing

The central part of the reconstruction algorithm is a SA MD protocol. During the annealing, a restraining potential keeps the center of mass of groups of AA particles close to their reference CG beads. The complete potential U^{tot} is described by eq. (1)

$$U^{\text{tot}} = U^{\text{AA}} + U^{\text{restr}}, \tag{1}$$

where U^{AA} represents the atomistic force field and U^{restr} is a harmonic potential,

$$U^{\text{restr}} = \sum_{i=1}^{n} \frac{k}{2}\left(\mathbf{r}_i^{\text{CG}} - \mathbf{r}_i^{\text{AA,com}}\right)^2. \tag{2}$$

In eq. (2), $\mathbf{r}_i^{\text{CG}}$ is the position of the reference CG bead i, $\mathbf{r}_i^{\text{AA,com}}$ the center of mass of the AA particles that are mapped to this bead, n is the total number of CG beads, and k a restraining force constant. The $\mathbf{r}_i^{\text{AA,com}}$ are updated at every simulation step.

Using this restraining potential, a SA MD procedure is used to generate low-energy structures. The temperature is gradually decreased from a high starting value to the desired target temperature, thus allowing the system to rapidly cross energy barriers and find a low-energy minimum at the end of the simulation. The crucial parameters are the starting temperature and the cooling rate; too low initial temperatures or too rapid cooling will not yield properly equilibrated final structures.

Because the AA particles are initially placed randomly around the corresponding CG beads, very large forces will occur at the beginning of the simulation and cause numerical instabilities. To ensure a stable simulation, these forces are reset to a specified threshold; this threshold is then linearly increased during the annealing simulation to increase the sampling. This is possible because the system optimizes in the course of the simulation, and thus large forces occur less frequently. The initial value for the force threshold, $F_{\text{cap},0} = 15,000\ \text{kJ mol}^{-1}\ \text{nm}^{-1}$, and its increase rate, $F_{\text{cap}} = F_{\text{cap},0} + At$, with $A = 100\,\text{kJ mol}^{-1}\,\text{nm}^{-1}\,\text{ps}^{-1}$, were chosen high enough to ensure that the relevant energy barriers are crossed, and at the same time low enough to avoid instabilities.

Release of the Restraints

At the end of the SA, the resulting atomistic structure is still coupled to the CG structure (cf. eq. (2)). However, in general, the (mapped) AA and CG minimum-energy structures will deviate because of the inherent differences between the force fields. Therefore, to release the strain on the AA structure, U^{restr} is smoothly removed at the end

of the reconstruction procedure to yield the final low-energy AA configuration.

The restrained SA simulation is carried out in the NVT ensemble, i.e., without pressure coupling. Simulating at constant volume prevents an expansion of the simulation box due to high initial forces. Any differences in the densities of the CG and atomistic systems can be taken into account at the end of the reconstruction by switching on pressure coupling after the release of the restraints. For CG systems simulated with the MARTINI force field, this difference is usually small.

Reconstruction of Solvent

A special situation arises for the reconstruction of solvents in which several small molecules are mapped to one CG bead, such as, e.g., water in the MARTINI model, where one CG water bead effectively represents four atomistic water molecules. To avoid water molecules diffusing away from their associated CG bead, a potential of the form

$$U_j^{\text{restr,W}} = \begin{cases} 0 & \text{for } r_{ij} \le r_{\text{CGW}} \\ \dfrac{k_{\text{W}}}{2}(r_{ij} - r_{\text{CGW}})^2 & \text{for } r_{ij} > r_{\text{CGW}} \end{cases} \quad (3)$$

is used in addition to the restraining potential described in eq. (2). In eq. (3), r_{ij} is the distance between the oxygen atom of AA water j and the center of mass of the four AA waters that belong to CG bead i, r_{CGW} the cutoff radius, and k_{W} the restraining force constant. Thus, every AA water molecule that moves out of the cutoff radius is driven back toward the center of mass. To ensure that the additional external force does not lead to a net center of mass movement of the four AA water molecules, a counter force is distributed among the remaining three water molecules such that $\sum_{j=1}^{4} F_j = 0$. This procedure is iteratively applied to every water molecule.

Implementation

We implemented the reconstruction algorithm into the program mdrun (v3.3), which is the main MD engine of the program package GROMACS.[5] The conversion between the AA and CG structures is done with the program g_fg2cg, which reads in both the AA and the CG topologies. Here, in the AA topology file, an additional mapping section is included that describes the correspondence between the two levels of resolution. With this information, g_fg2cg can either generate a CG structure from a given atomistic structure (in this case, the former is uniquely defined by the latter through the mapping) or generate an initial atomistic structure for the SA reconstruction simulation. The input parameters (cf., Table 1) for the restrained SA simulation are added to the MD-parameter (mdp) file and are read in by the GROMACS preprocessor (grompp). During the restrained SA simulation, the forces due to the external restraining potentials (Eqs. (2) and (3)) are computed at each integration step and are added to the forces derived from the original atomistic potential.

To simplify the mapping of large complex molecules, such as proteins or polycarbohydrates, we have modified the program pdb2gmx, which uses database files to generate a topology from a structure file. In these database files, now also the CG/AA mapping is defined for each building block, such as the individual amino acids.

Table 1. Mdp-Parameters for Reconstruction Simulation.

Parameter	mdp-option	Recommended value
Initial capping force $F_{\text{cap},0}$	cap_force	15,000 kJ mol^{-1} nm^{-1}
Capping increase rate A	cap_a	100 kJ mol^{-1} nm^{-1} ps^{-1}
Restraining force constant k	fc_restr	12,000 kJ mol^{-1} nm^{-2}
Radius of CG water r_{CGW}	r_CGW	0.21 nm
Water restraining force constant k_{W}	fc_restrW	400 kJ mol^{-1} nm^{-2}
Nr of steps to release restraints	rel_steps	5000
Annealing method	annealing	single
Annealing time	annealing_time	60 ps
Initial annealing temperature	annealing_temp	1300 K

Dihedral angle states that are separated by high energy barriers can lead to problems during the reconstruction, because the system might be trapped in the unwanted minimum. Such an unwanted minimum can be, for example, the cis rotamer of an amide bond in a polypeptide backbone. A possible solution is to add dihedral angle restraints to the topology. This can be done with the program g_dihfix, which processes the structure and topology files and selects those dihedral angles with a high energy barrier. The user can define a threshold for the dihedral angle force constant and thereby select certain dihedrals; each selected dihedral can then be restrained to a desirable value, for example, the one observed in the X-ray crystal structure of a protein. The code described in this work can be downloaded from our web page under http://md.chem.rug.nl/∼marrink/coarsegrain.html.

Simulation Details

All simulations were carried out using the GROMACS simulation package (version 3.3.1)[5] and the implementation of the reconstruction algorithm described in this work. In the CG simulations, version 2 of the MARTINI forcefield[8–10] was used together with a 20-fs integration time step. The other simulation parameters were set to the standard values described in the original publications.[9,10]

The atomistic simulations were carried out with a 2-fs integration time step, and the temperature was controlled by coupling to a Nośe–Hoover thermostat ($\tau_T = 0.1$ ps).[30] Because of the random initial placement of the atomistic particles (see Methods), no constraints were applied in the reconstruction simulations, except for SPC water. For the simulations of the NCQ tripeptide, the 53a6 parameter set of the GROMOS united atom force field[12] was used. For DPPC, we used the parameters published by Berger et al.[31] The force field parameters for cholesterol were taken from Ref. 32. In the vacuum simulations, no cutoffs for the nonbonded interactions were applied. The WALP20 peptide[33] was represented by the 43a2 parameter set of the GROMOS force field[11] and solvated with SPC water.[34] The system was simulated within periodic boundary conditions. Nonbonded interactions were calculated using a triple-range cutoff scheme: interactions within 0.9 nm were calculated at every time step from a pair list, which was updated every 20 fs. At these

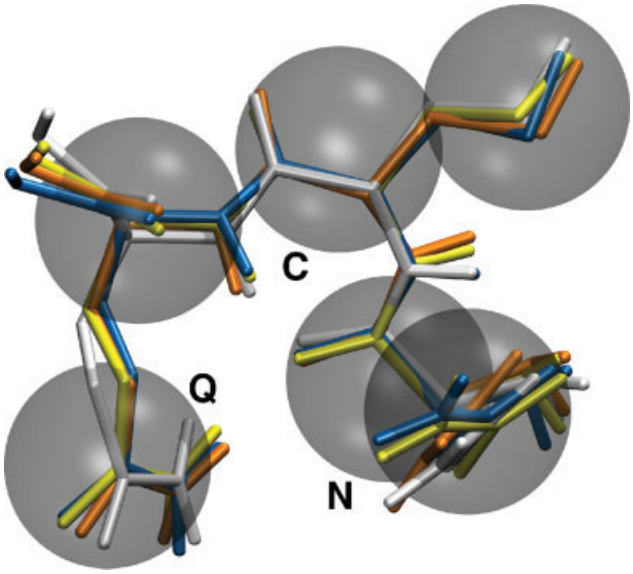

Figure 1. NCQ peptide. The coarse-grained structure is represented by gray spheres; an ensemble of four atomistic structures is shown as colored sticks. The N- and C-termini are capped with $-NH_2$ and $-COOH$ groups, respectively.

time steps, interactions between 0.9 and 1.5 nm were also calculated and kept constant between updates. A reaction-field contribution[35] was added to the electrostatic interactions beyond this long-range cutoff, with $\epsilon_r = 62$. In the simulations of bulk SPC water, a twin-range cutoff scheme was used, with a single cutoff at 0.9 nm and a pair list updated every 20 fs. Here, a long-range dispersion correction was applied in addition to the reaction field.

Example Applications

NCQ Peptide

As a first application of our algorithm, we reconstructed atomistic structures of an NCQ tripeptide in vacuum from CG structures. NCQ is shown in Figure 1 and is composed of the amino acids asparagine, cysteine, and glutamine. The application to NCQ has a twofold aim: (i) to find optimal parameters for the annealing procedure and (ii) to verify that proper ensembles are generated.

Parametrization of Annealing Procedure

Two sets of reconstruction simulations were carried out. The first set (Set1) was initialized from a structure taken from an equilibrium atomistic simulation at 300 K. To generate a reference CG structure for the definition of the restraints (see eq. (2)), the atomistic structure was converted to its CG representation using the MARTINI mapping. The second set of reconstruction simulations (Set2) started from a snapshot taken from an equilibrium MARTINI CG simulation at 300 K. For each combination of parameters, 1000 independent annealing simulations were carried out to generate an ensemble of reconstructed structures.

First, we investigated how the properties of the ensemble of reconstructed atomistic structures depend on two crucial parameters of the SA: the total annealing time, t_{tot}, and the initial temperature, T_{init}. In Figure 2A, the average potential energy of the ensemble of final structures of Set1 is plotted as a function of the length of the annealing simulation. As expected, E_{pot} decreases with t_{tot} and reaches a minimum at about -400.1 kJ mol^{-1} after 60 ps (standard deviation $= 20.8$ kJ mol^{-1}, std error $= 0.7$ kJ mol^{-1}). For comparison, the average E_{pot} obtained from a 400-ns equilibrium atomistic MD simulation is -392.8 kJ mol^{-1} (standard deviation $= 19.4$ kJ mol^{-1}, standard error $= 0.8$ kJ mol^{-1}). For short simulation times, the average energies are higher and the distributions are broader due to a number of high-energy structures within the ensemble. For longer simulation times, there are no such high-energy structures. The corresponding average energies are lower and the distributions are narrower, suggesting that the structures within the ensemble are similar to each other. Indeed, as shown in the inset, the rmsd with respect to the reference structure (i.e., the structure taken from the equilibrium atomistic simulation, see earlier) reaches its minimum value of 0.05 nm for simulation times longer than 60 ps. To obtain such a converged ensemble, the initial temperature T_{init} needs to be chosen high enough to overcome the relevant energy barriers during the simulation. Figure 2B shows that low-energy ensembles are obtained for $T_{init} \geq 1300$ K, which was thus chosen for the simulations shown in Figure 2A.

Figure 2C shows how the potential energy of the reconstructed ensemble depends on the force that couples the atomistic to the CG system. For Set1, the potential energy of the reconstructed ensemble depends only weakly on the restraining force constant (dashed line). Very weak force constants do not limit the sampling to only the lowest energy regions of the potential energy landscape, thus leading to a slightly higher E_{pot}. For Set2, i.e., the reference structure taken from the equilibrium CG simulation, a lower force constant yields ensembles with lower energies when compared with those obtained at higher force constants (solid line). This observed strain energy is due to the different regions of phase space sampled at the CG and AA levels of resolution. The larger this mismatch, the higher the strain. Thus, an atomistic system that is only weakly coupled to the CG system is allowed to sample lower energy configurations, which might be further away from the CG structure to which it is restrained.

On the one hand, the aim is to generate an atomistic ensemble that is close to its reference CG structure; on the other hand, strained structures should of course be avoided. This can be achieved in a two-step approach: first, an ensemble is generated using a rather high restraining force constant during the SA. Then, in a second step, the strain energy is released. Figure 2D shows the time evolution of the potential energy (solid line) and the rmsd with respect to the reference structure taken from the equilibrium CG simulation (dashed line). First, a 60-ps SA was carried out, followed by 20-ps equilibration at the final temperature of 300 K under the restraints. Then, after 80 ps, the restraining potential was removed within a time period of 10 ps. During this time, the system releases its strain energy in the fast degrees of freedom by relaxing to a local minimum (drop in E_{pot}), which is structurally further away from the reference structure (rise in rmsd). To release the additional strain energy in slow degrees of freedom or in degrees of freedom that involve high energy barriers requires more extensive sampling.

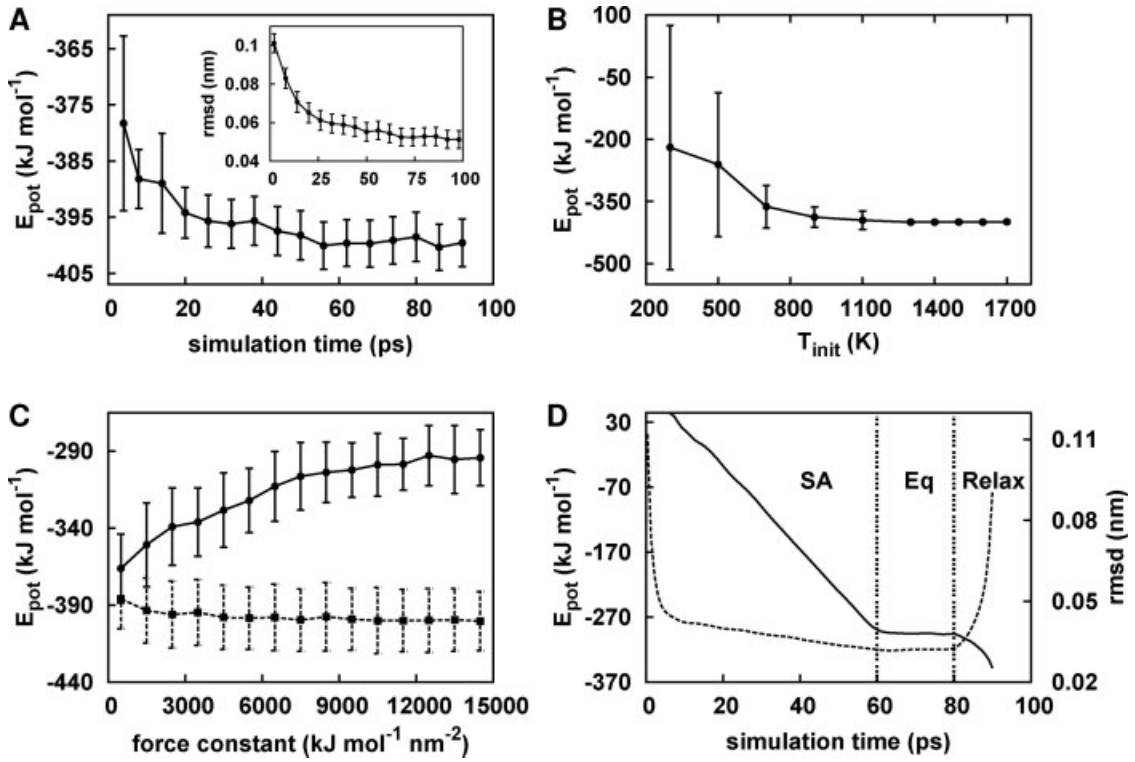

Figure 2. Reconstruction of NCQ peptide *in vacuo*. A: Average potential energy E_{pot} of ensemble of final structures of Set1 versus length of simulated annealing simulation, during which the temperature was decreased from 1300 to 300 K. Inset: all-atom rmsd of final structures with respect to reference structure. B: E_{pot} of final structures versus initial temperature T_{init} of simulated annealing simulation (80 ps each). Each data point in A and B represents the average of 1000 independent simulations; the standard deviation divided by 5 is plotted as error bars. C: Potential energy as a function of restraining force constant. Solid line: restraints defined relative to structure taken from equilibrium CG simulation (Set2). Dashed line: restraints defined relative to structure taken from equilibrium atomistic simulation (Set1). Each data point represents an average over 50 independent simulations; the standard deviation is shown as error bars. The starting temperature was 1300 K, and the simulation length 80 ps. D: Time evolution of E_{pot} (solid line) and rmsd with respect to reference structure (dashed line), averaged over 50 trajectories. Three phases of the reconstruction simulation are indicated by the dashed vertical lines: (i) simulated annealing (SA), (ii) constant temperature equilibration under the restraints, and (iii) relaxation of the system by removing the restraining potential.

Our results suggest that SA times larger than 60 ps and initial temperatures of 1300 K are sufficient to generate equilibrated atomistic structures from a CG NCQ peptide. In addition, possible differences between AA and CG force fields should be taken into account to genuinely choose the restraining force constant and to smoothly release the strain energy at the end of the reconstruction procedure.

Properties of Reconstructed Ensembles

Next, we more closely characterized the properties of the reconstructed atomistic ensembles of the NCQ peptide. In particular, we were interested in how much information at the atomistic level can be retrieved back from CG structures. To that end, we carefully analyzed the distributions of dihedral angles and E_{pot}, as well as their dependence on the mapping scheme and the starting structures used for the reconstruction. The different mapping schemes applied were an amino acid-to-1 mapping, i.e., each CG bead represents a complete amino acid; the MARTINI mapping, in which on an average four heavy atoms are mapped onto one CG bead; and a 2-to-1 mapping, i.e., each CG bead represents two heavy atoms. Additionally, also a 1-to-1 mapping was applied. In all simulations, an initial temperature of 1300 K was used, and the SA time was 80 ps. Figure 3 shows the distributions of the H$-$N$-$C$_\alpha-$C dihedral angle within asparagine (DihA, shown in the inset of Fig. 3C) and of the consecutive N$-$C$_\alpha-$C$-$N dihedral angle (DihB, shown in the inset of Fig. 3D) as two representative examples; similar results were obtained for other dihedral angles (not shown).

Figure 3A shows the distributions of DihA obtained from 1000 reconstruction simulations initialized from a single AA structure (Set1, see earlier). The black dotted line indicates DihA in the reference structure, which is at about $-80°$. The coarse amino acid-to-1 mapping yields an ensemble in which three dihedral states are populated (magenta curve): the reference state at DihA $= -60°$, the next

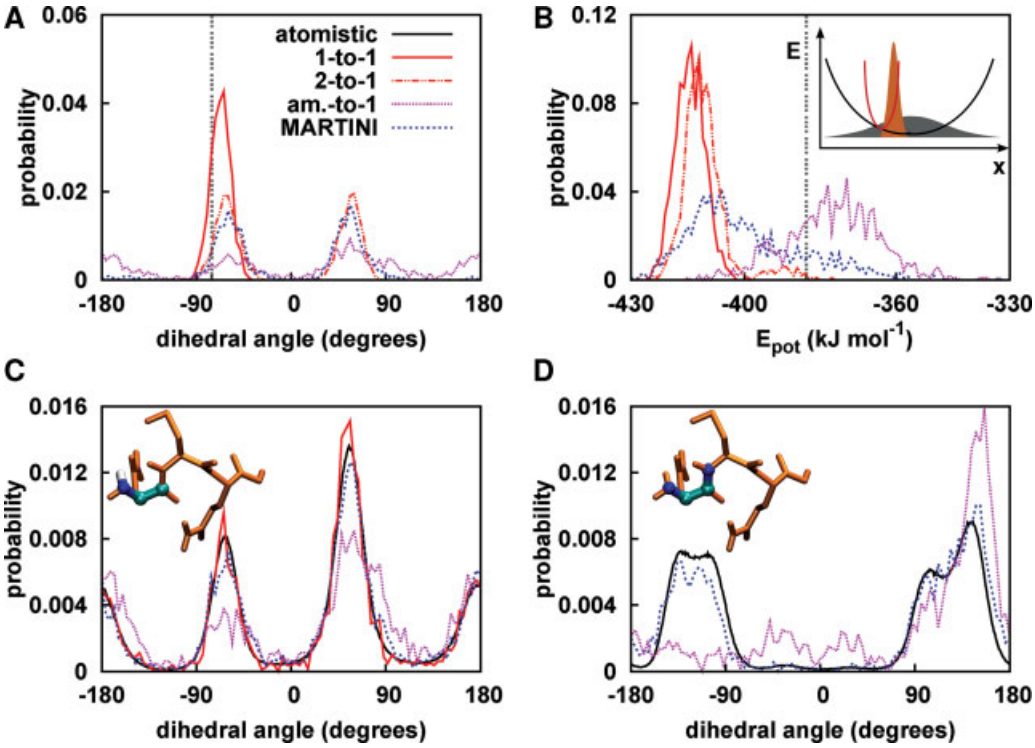

Figure 3. Dihedral angle and potential energy distributions of NCQ, obtained using various mapping schemes. Black solid line: data obtained from atomistic equilibrium simulation; red solid line: 1-to-1 mapping; orange dot-dashed line: 2-to-1 mapping; blue dotted line: MARTINI mapping; magenta dotted line: amino acid-to-1 mapping. The dashed gray line indicates the reference structure. A: Distributions of dihedral angle DihA (C, inset) within ensemble of 1000 structures; all reconstructions were started from a single structure taken from an equilibrium atomistic simulation. B: Potential energy distributions for the ensembles shown in A. Inset: distributions in a local potential minimum (black line) with and without restraining potential (gray and orange distributions, respectively; the red line depicts U^{restr}) are schematically depicted. C: Distributions of DihA (inset); 1000 reconstructions were initialized from equally spaced snapshots taken from a 200-ns equilibrium atomistic simulation. D: Distributions of DihB (inset); initial structures the same as in C.

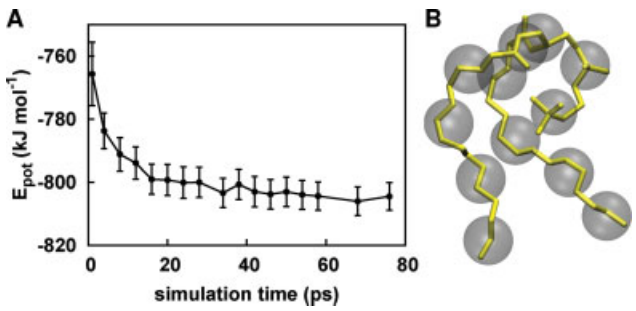

Figure 4. Reconstruction of a DPPC molecule *in vacuo*. A: Average potential energy E_{pot} of final structures versus length of simulated annealing simulation (T_{init} = 1300 K). For each annealing time, 1000 independent simulations were carried out; error bars represent the standard deviation divided by 5. B: MARTINI (gray spheres) and atomistic (yellow sticks) representations of DPPC.

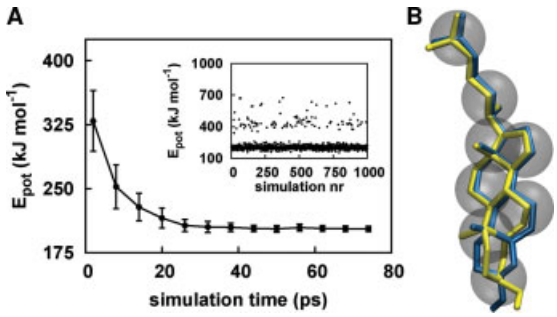

Figure 5. Reconstruction of a cholesterol molecule *in vacuo*. A: Average potential energy E_{pot} of final structures versus length of simulated annealing simulation (T_{init} = 1300 K). For each annealing time, 1000 independent simulations were carried out; error bars represent the standard deviation divided by 5. Inset: the potential energy of the final structure obtained after 14 ps of simulated annealing is plotted for each individual simulation. B: MARTINI (gray spheres) and atomistic (yellow and blue sticks) representations of cholesterol. The blue structure is correct, whereas the yellow structure has an inverted stereocenter at C10.

minimum at +60°, and a third state at 180°. In contrast, with the MARTINI mapping (blue dotted curve) and the 2-to-1 mapping (red dot-dashed curve), only the two dihedral states at +/−60 degrees are populated. As expected, with the 1-to-1 mapping (red solid curve), only the reference dihedral minimum is populated, because the restraining does not allow for the population of a different state in this case.

In Figure 3B, the distributions of E_{pot} obtained from Set1 are shown for the different mapping schemes. The reference potential energy of the single structure taken from the equilibrium AA simulation is plotted as a dashed gray line at -380 kJ mol^{-1}. The distribution obtained using the amino acid-to-1 mapping (magenta curve) is centered at -370 kJ mol^{-1}, which is higher than the reference value. MARTINI mapping yields a distribution with an average potential energy of -400 kJ mol^{-1} (blue curve). This is lower than the reference energy of -380 kJ mol^{-1}, which, however, falls into the tail of the distribution. The distributions obtained with the 2-to-1 (dashed red curve) and 1-to-1 mapping (solid red curve) have even lower average energies and are very sharp. Here, the restraining potential limits the accessible phase space to a confined region that has a lower average energy when compared with the one sampled in a free simulation, as schematically illustrated in Figure 3B, inset.

The question is whether the observed population of the dihedral states is correct, i.e., represents the equilibrated ensemble for the AA force field, or might be an artifact of the reconstruction procedure. To address this question, a 400-ns atomistic equilibrium simulation was carried out at 300 K. The dihedral angle distributions obtained from the two 200-ns halves of this trajectory are virtually identical (data not shown), indicating that the simulation is converged in this respect. The black solid line in Figure 3C shows that all three dihedral states are populated. Here, the dihedral state at +60° has the highest population. We initialized 1000 reconstructions from equally spaced snapshots taken from the first half of the 400-ns atomistic equilibrium trajectory. The distributions shown in Figure 3C show that all three dihedral states are populated, irrespective of the mapping. The relative populations of these three states are in agreement with the atomistic equilibrium simulation.

As described earlier, we have optimized the reconstruction procedure for the MARTINI mapping and found that it also works well for finer mapping schemes. However, a coarser mapping (i.e., a lower number of restraints) could lead to artifacts, because the system is allowed to sample high-energy regions in configurational space, where it can become trapped. This is observed for the reconstructions of NCQ using the amino acid-to-1 mapping, which yields an ensemble with a high E_{pot} (see magenta dotted curve in Fig. 3B). This pitfall becomes also evident from an analysis of DihB (Fig. 3D), where the amino acid-to-1 mapping fails to capture the minimum at −120° and strongly overpopulates the dihedral minima at +/−30°. A slower annealing procedure might rescue the problem; however, we do not investigate this issue in more detail here.

To summarize, the reconstructed atomistic ensembles correctly represent the possible dihedral states and energy minima. The reconstruction procedure does not pose any bias on the obtained distributions and yields proper ensembles of atomistic structures.

DPPC and Cholesterol

Our next goal was to investigate whether and how the parameters of the reconstruction procedure that were optimized for the NCQ tripeptide can be transferred to other molecules as well. To this end, we chose DPPC and cholesterol as two typical examples for lipids and sterols, respectively. The reference structures for the definition of the restraints during the reconstruction simulations were taken from atomistic equilibrium MD simulations of the single molecules in vacuum; the CG representations were constructed via MARTINI mapping.

The potential energies of the reconstructed ensembles of DPPC and cholesterol are shown in Figures 4 and 5, respectively. E_{pot} drops with increasing simulation time, similar to NCQ (cf., Fig. 2A). However, the energy plateau is reached already after about 30 ps of SA, which is about two times faster when compared with NCQ. This difference is expected, because the energy barriers that have to be overcome during annealing of DPPC and cholesterol are lower, *inter alia*, because of the absence of hydrogen bonds.

A special situation arises for cholesterol: it has two stereocenters at C10 and C13, where the chiral carbon atom is connected to four different neighboring groups (Fig. 5B). Because of the random initial placement of the atoms close to their reference CG beads (see Methods), there is a chance that, e.g., the methyl group bound to C10 is initially on the "wrong" side of the plane spanned by the other three bound groups. This methyl group would have to tunnel through this plane to end up on the correct side and thus overcome a high energy barrier. Thus, although this process is possible during the reconstruction due to the applied force capping, for too fast annealing, one (or both) of the stereocenters can be trapped in the wrong configuration, as shown in Figure 5B, yellow structure. For a short annealing time of 14 ps, the final ensembles can be subdivided into three populations (Fig. 5A, inset): in most of the cases, correct structures with a low potential energy were obtained. In about 8% of the cases, a structure was obtained in which one of the two stereocenters is inverted, whereas a structure with both stereocenters inverted occurred at a probability of only 2%. For annealing times longer than 30 ps, no such incorrect structures were generated.

There are other means to tackle the problem of inverted stereocenters. For example, we applied soft-core interactions, which basically introduces two different force capping thresholds for the bonded and the nonbonded forces, because only the latter are affected by the soft-core interactions.. By this means, a proper ensemble of reconstructed cholesterol molecules was generated even within shorter annealing times and from lower initial temperatures (data not shown). Other authors have suggested similar reconstruction procedures, where initially only the bonded interactions are used and the nonbonded interactions are gradually introduced during the course of the reconstruction.[22,36,37]

To summarize, we found that the annealing parameters optimized for NCQ can be transferred to DPPC and cholesterol as two representative examples for lipid and sterol molecules, respectively. For these molecules, annealing times of about 30 ps are sufficient, which is about a factor of two faster when compared with the NCQ peptide.

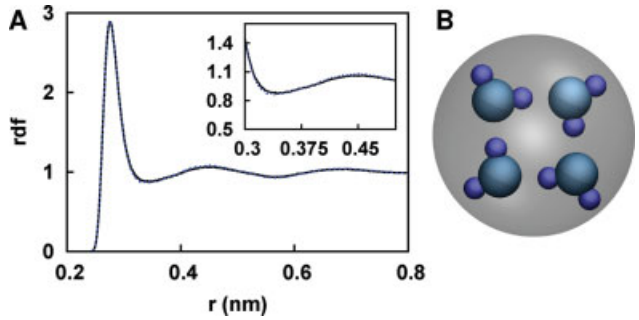

Figure 6. A: The oxygen-oxygen radial distribution functions of restrained (dashed blue line) and free SPC water (solid black line) overlap. A zoom on the central region of the RDF is shown in the inset. B: Four SPC water molecules (blue) are restrained to one MARTINI water bead (gray sphere). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

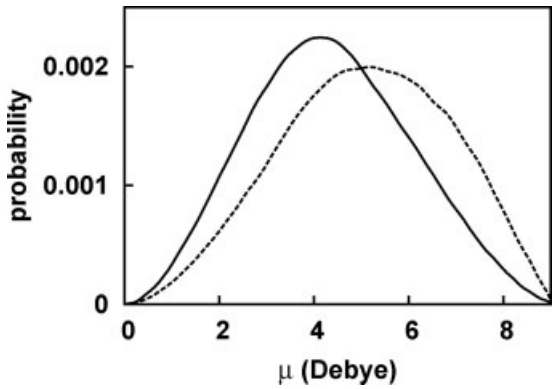

Figure 7. Comparison between dipole moment distributions of restrained (dashed line) and free SPC water (solid line), analyzed in groups of four water molecules.

Water

Typical biomolecular simulation systems contain large amounts of water; thus, the generation of an atomistic water configuration from a CG representation is an important step to investigate. We have reconstructed an SPC water box starting from a pre-equilibrated box of 400 MARTINI water beads, i.e., 1600 SPC water molecules at 300 K. The systems were simulated within the NVT ensemble, and the box size was 48.4 nm^3, corresponding to a density of 988.7 kg m^{-3}. The reconstruction simulation was started at an initial temperature of 400 K, and the annealing time was 60 ps. A CG water radius $r_{CGW} = 0.21$ nm and a water restraining force constant $k_W = 400$ kJ mol^{-1} nm^{-2} were used (see Table 1 and eq. (3)). The restraining potential was kept on during the entire simulation. For analysis, the reconstruction simulation in the constraint ensemble was extended to 100 ns at 300 K. For comparison, a 20-ns NVT simulation of the same SPC water box without restraints was carried out.

In Figure 6, the oxygen–oxygen radial distribution functions (RDFs) obtained from the restrained and free SPC water simulations are compared. Both RDFs coincide very well with each other, showing that the restraining to the CG system does not perturb

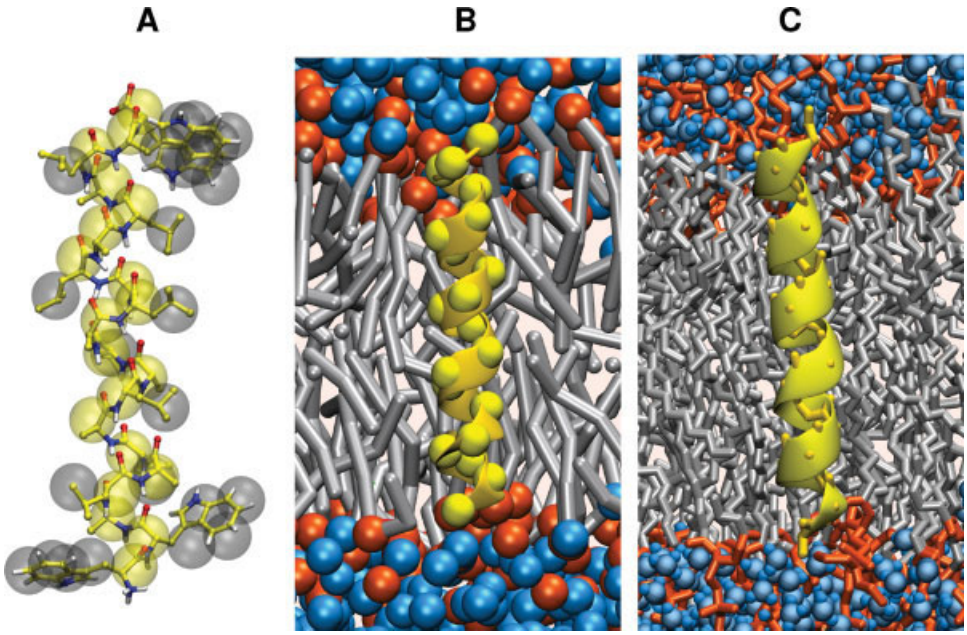

Figure 8. WALP20 transmembrane peptide in a solvated DPPC bilayer. A: Atomistic (ball and sticks) and MARTINI (gray and yellow spheres) representations of WALP20. B: CG (MARTINI) representation of the system. The WALP peptide is shown as a yellow α-helix; lipid tails and head groups as gray sticks and red spheres, respectively; water molecules are colored blue. C: Reconstructed atomistic system, obtained after 80 ps of restrained simulated annealing. Color scheme same as in B. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the local packing of water. However, as shown in Figure 7, the dipole moment distributions of the two systems are different: the correlation between individual dipoles of water molecules that are grouped together by the restraining potential yields a dipole distribution (dashed line) that is slightly shifted with respect to the one of free SPC (solid line). The average dipole moment of four restrained water molecules is 5.1 Debye, whereas it is 4.3 Debye for free SPC. As a consequence, the dielectric constant of the restrained system is $\epsilon_r = 72$, which is slightly higher than $\epsilon_r = 65$ for free SPC.[34,38] These values can be compared to the experimental value of $\epsilon_r = 78$ at 298 K.

Our results show that the structural properties of reconstructed atomistic water that is still restrained to the CG system correspond to those obtained from free atomistic simulations, and that the dielectric constant of the reconstructed water is in between that of free SPC and experiment. Similar results were obtained from reconstruction simulations that were run at a constant final temperature of 300 K (results not shown), because of the low energy barriers involved in the reorganization of water hydrogen bonds. Note that the simulations were carried out in the NVT ensemble, i.e., at constant box volume. Applying the reconstruction approach within an NPT simulation would yield a too condensed system due to a negative contribution to the overall pressure from the additional restraining potential that keeps the water molecules together. This can be remedied by a reparameterization of the atomistic water model used, if required.[39]

WALP Peptide in Solvated DPPC Bilayer

In the previous sections, we have tested and optimized a set of parameters for the restrained annealing simulations, which turned out to be well suited for a number of different biomolecules in the gas phase and for bulk water. Finally, our aim is to demonstrate that our reconstruction procedure can also be applied to a real-life problem, i.e., a condensed-phase biomolecular system with several components. Here, we chose the WALP20 peptide embedded in a solvated DPPC bilayer as a typical application.

The WALP20 peptide (Fig. 8A) is an α-helical transmembrane peptide composed of a hydrophobic central repeat of alanine and leucine residues, flanked by two tryptophans ($W_2-(AL)_8-W_2$).[33] To generate a starting structure for the subsequent reconstruction simulations, a 40-ns CG simulation (MARTINI force field) of a system containing a WALP20 peptide embedded in a membrane bilayer composed of 112 DPPC lipids, solvated by 1186 water beads, was carried out at $T = 323$ K in the NPT ensemble ($p = 1$ bar). During this simulation, the helical structure of the peptide was maintained by means of additional dihedral restraints.[10]

For each annealing time, 10 reconstruction simulations were started from the above structure, during which the temperature was linearly decreased from 1300 to 323 K. In the reconstruction simulations, the amide bonds were restrained to the trans conformation (see Implementation). Figure 8 shows a part of the simulation system before (B) and after (C) the reconstruction. In Figure 9, the helicity of the reconstructed WALP20 peptides is plotted as a function of the annealing time. For simulation times longer than 20 ps, the helicity reaches a plateau at about 17 residues. The same value was obtained from a free 10-ns atomistic simulation of the same system. Although the helical structure of the peptide is already fully formed in about 20 ps, the potential energy of the system (dashed line in Fig. 9) indicates that also in this case, simulation times longer than about 60 ps are required to obtain a low-energy structure.

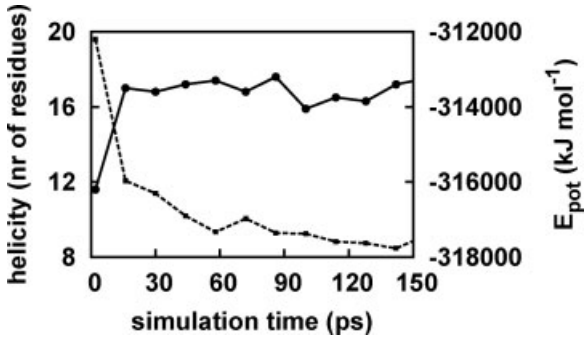

Figure 9. Helicity (solid line) of WALP20 peptide and potential energy (dotted line) of simulation system as a function of annealing time. The helicity was obtained by means of the DSSP program.[40] Each data point represents an average over 10 reconstruction simulations; the potential energy was evaluated for the last snapshot of each simulation.

The reconstruction of a large system composed of a WALP peptide embedded into an explicitly solvated DPPC bilayer confirms that the method can also be successfully applied to larger systems. Another successful reconstruction was performed for a number of different lipid membranes in the presence of the antimicrobial peptide magainin, as described elsewhere.[41] It might at first sight seem surprising that it is possible to apply the reconstruction procedure that was optimized for single molecules *in vacuo* to a condensed-phase system without further adjustments to the parameters. However, the initial placement of the atomistic particles close to their reference CG beads limits the reconstruction to a local sampling problem, and the interactions with the surrounding atoms do not significantly slow down the reconstruction.

Summary and Conclusions

In this work, we describe the implementation of an algorithm to reconstruct atomistic details from CG structures into GROMACS, a free and highly efficient MD program package. A three-step approach is used to optimize low-energy ensembles: First, the atomistic particles are positioned close to their reference CG beads. Second, a SA MD procedure is used, during which the atomistic and CG systems are coupled via restraints. Third, the coupling is gradually removed to ensure a smooth relaxation of the reconstructed atomistic ensemble.

We systematically tested and optimized the reconstruction procedure for single molecules in the gas phase, such as the NCQ tripeptide, the DPPC lipid, and cholesterol. In addition, we reconstructed bulk atomistic water from CG water. These reconstructions yielded proper low-energy atomistic ensembles underlying the CG representations. Finally, a system composed of a WALP20 transmembrane peptide embedded in a solvated DPPC bilayer was transformed from a CG to an atomistic representation. The peptide adopts its α-helical secondary structure when reconstructed within the membrane bilayer. In these examples, the MARTINI CG force

field was used in combination with the GROMOS96 atomistic force field. However, because it does not rely on libraries of predefined fragments, our reconstruction algorithm can be straightforwardly applied to a wide range of force fields and molecule types.

From the applications presented in this work, we conclude that, to be on the safe side, annealing times of about 60–80 ps in conjunction with initial temperatures of 1300 K should be used for obtaining low-energy ensembles of reconstructed structures. These simulation times may be considered long, but may be speeded up by applying nonlinear temperature annealing schemes. Additionally, because the annealing is merely used to optimize low-energy structures, and not to calculate dynamic properties of the system, the mass of the hydrogen atoms can be increased, which would allow for a larger integration time step.

Because of the loss of information inherent to transforming an atomistic to a CG structure, the reverse transformation cannot be expected to always yield a single atomistic structure, but rather an ensemble of structures that all correspond to the same CG structure and occur according to their Boltzmann probabilities. Thus, as demonstrated in this work, it is important that the reconstruction is carried out several times to generate such an ensemble.

By unraveling the atomistic details underlying a CG structure, the reconstruction algorithm allows to switch between the different levels of resolution. In some applications, it might, for example, be useful to temporarily switch from a CG to an atomistic representation to investigate special events. By this means, the reconstruction method allows to check and validate the results and predictions obtained with CG models against atomistic models, thereby combining the efficiency of the former with the accuracy of the latter.

Appendix: List of Commands to Run Reconstruction Simulation

1. Topology preparation for atomistic system. Here, aa.gro is an atomistic structure, for example, of the NCQ peptide. The [mapping] section will be included in the output topology aa.top.
   - pdb2gmx -ignh -missing -f aa.gro -p aa.top
2. Definition of dihedral angle restraints for certain dihedrals within the peptide. Here, every dihedral with a force constant larger than 10 kJ mol^{-1} (-fc option) is listed in the (dihedral_restraints) section in the new topology aa_restr.top.
   - g_dihfix -c aa.gro -p aa.top -fc 10 -o aa_restr.top
3. Initial placement of atomistic particles around the corresponding CG beads. When option *n* is set to 1 and an atomistic structure is used as an input, the program calculates the positions of the CG beads from the positions of the underlying atomistic particles. When *n* is 0 and a CG input structure cg.gro is provided, a random atomistic structure aa_random.gro is generated. Note that both CG and AA topologies (cg.top and aa_restr.top, respectively) have to be provided.
   - g_fg2cg -pcg cg.top -pfg aa_restr.top -c cg.gro -n 0 -o aa_random.gro
4. Restrained simulated annealing simulation using aa_random.gro as the input structure. In the sa.mdp file, the parameters from Table 1 should be included. Note that if dihedral restraints are used, the user has to additionally define "dihre = simple" in the mdp file and specify the dihedral restraints force constant (e.g., via "dihre_fc = 1000").
   - grompp -p aa_restr.top -f sa.mdp -c aa_random.gro
   - mdrun -coarse cg.gro

References

1. van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glaettli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. Angew Chem 2006, 45, 4064.
2. Karplus, M.; McCammon, J. A. Nat Struct Biol 2002, 9, 646.
3. Marrink, S. J.; de Vries, A. H.; Tieleman, D. P. BBA Biomembr 2009, 1788, 149.
4. Voth, G. A., Ed. Coarse-Graining of Condensed Phase and Biomolecular Systems; CRC-Press, Boca Raton, 2008.
5. van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. J Comput Chem 2005, 26, 1701.
6. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. J Chem Theory Comput 2008, 4, 435.
7. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Science 1983, 220, 671.
8. Marrink, S. J.; de Vries, A. H.; Mark, A. E. J Phys Chem B 2004, 108, 750.
9. Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. J Phys Chem B 2007, 27, 7812.
10. Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. J Chem Theory Comput 2008, 4, 819.
11. van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. The GROMOS96 Manual and User Guide Zürich, Switzerland, 1996.
12. Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. J Comput Chem 2004, 25, 1656.
13. Lyman, E.; Zuckerman, D. M. J Chem Theory Comput 2006, 2, 656.
14. Christen, M.; van Gunsteren, W. F. J Chem Phys 2006, 124, 154106.
15. Liu, P.; Shi, Q.; Lyman, E.; Voth, G. A. J Chem Phys 2008, 129, 114103.
16. Praprotnik, M.; Delle Site, L.; Kremer, K. J Chem Phys 2005, 123, 224106.
17. Praprotnik, M.; Delle Site, L.; Kremer, K. Phys Rev E 2006, 73, 066701.
18. Praprotnik, M.; Delle Site, L.; Kremer, K. Annu Rev Phys Chem 2008, 59, 545.
19. Ensing, B.; Nielsen, S. O.; Moore, P. B.; Klein, M. L.; Parrinello, M. J Chem Theory Comput 2007, 3, 1100.
20. Heyden, A.; Truhlar, D. G. J Chem Theory Comput 2008, 4, 217.
21. Milano, G.; Müller-Plathe, F. J Phys Chem B 2005, 109, 18609.
22. Hess, B.; Leon, S.; van der Vegt, N. F. A.; Kremer, K. Soft Matter 2006, 2, 409.
23. Harmandaris, V. A.; Adhikari, N. P.; van der Vegt, N. F. A.; Kremer, K. Macromolecules 2006, 39, 6708.
24. Shih, A. Y.; Freddolino, P. L.; Sligar, S. G.; Schulten, K. Nano Lett 2007, 7, 1692.
25. Heath, A. P.; Kavraki, L. E.; Clementi, C. Proteins: Struct Funct Bioinformatics 2007, 68, 646.
26. Carpenter, T.; Bond, P. J.; Khalid, S.; Sansom, M. S. P. Biophys J 2008, 95, 3790.
27. Villa, A.; Peter, C.; van der Vegt, N. F. A. Phys Chem Chem Phys 2009, 11, 2077.
28. Havranek, J. J.; Baker, D. Protein Sci 2009, 18, 1293.
29. Correa, P. E. Proteins 1990, 7, 366.

30. Hoover, W. G. Phys Rev A 1985, 31, 1695.
31. Berger, O.; Edholm, O.; Jähnig, F. Biophys J 1997, 72, 2002.
32. Marrink, S. J.; de Vries, A. H.; Harroun, T. A.; Katsaras, J.; Wassall, S. R. J Am Chem Soc 2008, 130, 10.
33. de Planque, M. R. R.; Killian, J. A. Mol Membr Biol 2003, 20, 271.
34. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Intermolecular Forces; D. Reidel Publishing Company: Dorodrecht, 1981.
35. Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. J Chem Phys 1995, 102, 5451.
36. Spyriouni, T.; Tzoumanekas, C.; Theodoru, D.; Müller-Plathe, F.; Milano, G. Macromolecules 2007, 40, 3876.
37. Brünger, A. T.; Karplus, M. Acc Chem Res 1991, 24, 54.
38. van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. J Chem Phys 1998, 108, 10220.
39. Fuhrmans, M.; Sanders, B. P.; Marrink, S. J.; de Vries, A. H. Theor Chem Acc DOI: 10.1039/6901615e 2009.
40. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577.
41. Rzepiela, A. J.; Sengupta, D.; Goga, N.; Marrink, S. J. Faraday Discuss DOI: 10.1007/s00214-009-0590-4.